Analysis of Delay in Mathematical Switching Models for Data Systems

By D. G. HAENSCHKE

(Manuscript received August 17, 1962)

Traffic delay, caused by temporary all-lines-busy conditions, is analyzed for three mathematical switching models. They are classified as "address camp-on," "retrial," and "message storage" models. The models are designed to permit a study of basic traffic theoretical problems encountered in the rapidly growing field of data communications, but they are not identical with any of the existing data switching systems. Each model assumes that a message is switched only through one switching center which must establish connections via line groups to one or more addressed receiving stations, i.e., each model contains only a single switching center. Numerical results for the average delay on all messages are obtained on the IBM 7090 computer.

I. INTRODUCTION

Switching centers can be used to link together communication lines for the transmission of data between a variety of business machines and computers. Due to randomness in the required interconnections a switching center may occasionally not find an idle line to a particular receiving station, so that a delay can occur. More than one method can be followed when a switching center finds all lines to a receiving station busy. Some switching models appear to obtain lines to the addressed receiving stations in a shorter time than other switching models. This means that with one switching model a given delay requirement can be met with fewer lines than with another switching model. This is not to say that the model which would render a given grade of service with the least number of lines also is the most desirable from an economic point of view, because delay is only one factor which enters into the choice between data switching systems. Components of a switching system, such

as lines, memory, etc. do not bear the same price tag, and minimizing the number of components of one kind does not ensure economical efficiency.

Interest in the particular traffic engineering problems of data switching has been present for at least 15 years. Yet, traffic engineering work was mainly concentrated on classical telephone trunking problems, and a variety of such fundamental problems have been worked out. Some of the data switching systems which are being studied or are now in use cannot be analyzed by standard mathematical approaches of traffic theory because the operating conditions differ from those of the mathematical models used in the analysis of classical telephone problems. The understanding of the fundamental traffic theoretical problems encountered in data switching is a prerequisite for an exact mathematical analysis of message delay in data systems. The fundamental problems need to be studied on simplified models which lend themselves best to mathematical treatment and, therefore, will not be identical with any of the present data systems in use. We have constructed for study three hypothetical models which we call "address camp-on," "retrial," and "message storage" models, and have analyzed message delay for each of them.

Message delay is defined as the delay between initial request by the switching center for a line, and the moment the message is released from the switching center for transmission. The switching center handles messages in a manner described by one of the three switching models. The delay is caused by temporary all-lines-busy conditions in the line groups which connect the switching center with the addressed receiving stations. This type of traffic delay must not be confused with the total delay from the time a message is ready at the data source and the time the message is actually received at a destination. No account is taken of messages which are switched through more than one switching center in tandem.

This study, then, shall not be looked upon as an attempt to make a choice between switching systems, since such a choice cannot be based solely on the delay performance of mathematical models. A true comparison between switching systems must include other factors, as for instance the cost of memory and logic, loading of transmitters, and loading of incoming lines, all of which are neglected here.

II. DESCRIPTION OF MATHEMATICAL SWITCHING MODELS

The following describes each of the three switching models. The description is preceded by an outline of features which are common to each model. The mathematical derivations given in the appendices and the delay curves are based on these models.

2.1 *Common Features*

Think of a data source feeding messages into a switching center that has a large number of line groups radiating from it (see Fig. 1).

Each line group connects one, and only one, receiving station with the switching center. A receiving station is capable of receiving from all lines in its group simultaneously. There might be one or more lines per group, but each group contains the same number of lines c. Full access is given to each line in a group. A message is said to have A addresses when a copy of the message must be transmitted over A different line groups to A different receiving stations. The number of addresses per message remains constant for all messages. The switching center is responsible for transmitting a copy of the message to each of the addressed receiving stations. The addresses of a message are chosen at random from a large number of possible receiving stations. This permits us to assume that all line groups are independent of each other. Messages are originated and addressed in such a way that a_I, the information load offered to a group, is the same for every group. The information load is defined as the number of first, i.e., unrepeated, message attempts which are expected to be generated during an interval equal to one average message length. First attempts are made Poisson distributed in time, meaning that the probability that exactly k first attempts are generated during an interval of length t is given by

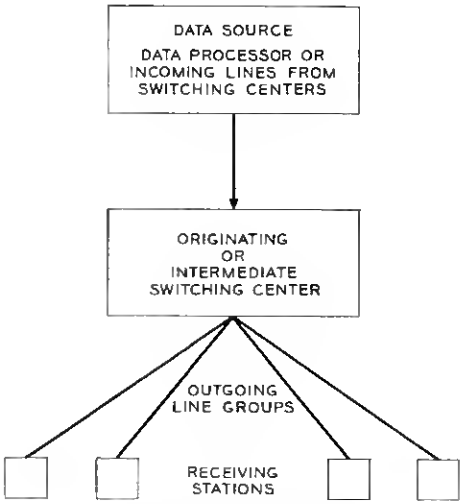

Fig. 1 — Switching model.

$$\Pr\ (k,t) = \frac{e^{-a_I t}(a_I t)^k}{k!},$$

in which t is in units of the average message length. The length of all messages is exponentially distributed with mean 1. In the mathematical derivations the average message length is taken as the unit of time. An exponential message length distribution is chosen, since it is believed that it will serve as a good approximation in a larger number of practical cases than a constant length. The instant a new or repeated message is originated, the switching mechanism begins to hunt for an idle line to each of the addressed receiving stations. No delay is imposed by the switching mechanism itself. Each new message is eventually delivered to the respective receiving stations, i.e., no messages are lost. The system is in statistical equilibrium, which is to say that the system is in the steady state such that the average number of messages in the system during any long interval of time remains constant.

When all lines in a group to one or more of the receiving stations are busy we must find a way of delaying delivery. The camp-on and storage models assume that blocked requests form queues at the switching center. The retrial model assumes that blocked requests are withdrawn from the switching center and reoffered at a later time.

The reader has doubtlessly observed the very simplified and idealized set of common features on which the switching models are based to permit mathematical analysis. The same applies to the features which are unique to each of the three models.

2.2 *Address Camp-On Model*

When a line group to an addressed receiving station is blocked, the request for service in this group will camp-on and wait in the order of arrival until a line is assigned by the mechanism which scans continuously for idle lines. The assignment of available lines to waiting requests is done on a "first come, first served" basis. When a line is assigned it is immediately made busy. The message, however, is not released from the data source until lines to all addressed receiving stations are secured. When the last line is obtained, the message content is released and transmitted simultaneously to each of the addressed receiving stations, after which the lines are released. The holding time of a line in the camp-on model is made up of the sum of two random variables: namely, the exponentially distributed message length and the time spent waiting until lines are secured to all addressed receiving stations.

2.3 *Retrial Model*

In the retrial model, a message is released from the data source only when lines are found at the switching center to all addressed receiving stations. If even one line group is blocked, the message is not transmitted to any of the addressed receiving stations and is temporarily cleared from the switching mechanism without making any lines busy. A blocked message is reoffered any number of times from the data source after a constant time interval τ, until an idle line is found simultaneously to each addressed receiving station. At a retrial of a blocked message, an attempt to seize an idle line is made in the same groups as at the previous attempt. The message delay is determined by the number of attempts made and by the length of the constant retrial interval τ. The holding time of a line in the retrial model is equal to the exponentially distributed message length.

Another way of making retrials is to let the delay in the delivery of the message content to any one addressed station be independent of the delay to the other addresses of a multiaddress message. In this case, a message having A addresses would be considered to consist of A one-address messages and the delay would be that given for the one-address case of the retrial model.

2.4 *Message Storage Model*

Message storage is analysed on a model in which requests for lines in a busy group form queues in the order of arrival. In the multiaddress case, some addresses of the message may find their line groups busy while other addresses may obtain lines to the addressed stations with no delay. The model assumes that the message is released with no delay to stations which are not blocked, and that the delivery to a blocked station is delayed only until the instant a line is found by the switching mechanism which scans continuously for idle lines. As in the retrial model, the line holding time is equal to the exponentially distributed message length.

III. METHOD OF ANALYSIS

The delay performance is analysed as messages are switched through one switching center which employs one of the three described switching models. It should be pointed out that the results obtained here apply only to the mathematical models used. All approximations mentioned in the analysis are approximations of the model to which they refer.

The mathematical analysis of the delay performance of the address camp-on model is given in Appendix A. The problem is to find the total occupancy of the outgoing line groups. In the multiaddress case, outgoing lines can be held busy in excess of the message length. This excess holding time increases the load carried from the useful information load a_I to a total load a_T. The excess holding time is the average length of time between line seizure and the time lines are found for all addresses of the message. Erlang delay probability is used by the introduction of an approximation which assumes that the total holding time of an outgoing line is exponentially distributed. No explicit expression is derived for the total load a_T. Solutions for a_T are found by solving (7) and (10) of Appendix A in an iterative computer program. The average delay follows from (11).

In Appendix B the mathematical analysis is given for the retrial model. The retrial method has been under consideration for application in both military and commercial data systems, and this method is also used in voice telephone communications. The mathematical analysis of delay in systems in which blocked attempts are reoffered is one of the fundamental traffic problems for which an exact solution is not available. The prospect of using the retrial method in data systems emphasizes the need to treat such systems analytically. The analysis given here is not exact because a number of approximations had to be introduced to obtain numerical results. Since the retrial method is a basically unsolved problem it must first be studied in its simplest form, which exists for the case of one address per message. Considerable effort, therefore, is spent in Appendix B on the discussion of the one-address case. Our approach to the retrial problem is to find approximations for the unconditional state probability of finding i lines in a group of c lines busy, $0 \leqq i \leqq c$. Then, approximations are found for the conditional probabilities of finding i lines busy at $t_0 + \tau$, when the state of the group is known at t_0, $t_0 - \tau$, $t_0 - 2\tau$, etc. The delay for the one-address case follows from (23) of Appendix B. For the three-address case the delay is computed from (31) in a Markov process which is in itself an approximation of the retrial problem since it accounts only for a first-order dependency.

The basic problem in the retrial model is to find approximations of the conditional probabilities mentioned above. These are obtained by integrating a set of differential equations (16), using a line request rate $\omega(t)$ which by itself is conditioned on previous states of the line group and, therefore, is dependent on time. The line request rate $\omega(t)$ appears as a coefficient in (16). Since $\omega(t)$ can be expressed only as a function of

solutions to (16), we cannot find $\omega(t)$ explicitly, but must compute it in a long process of progressive iterations. It will become apparent from Appendix B that not all approximations made can be clearly justified, but the results obtained are sufficiently accurate for comparison with other switching models. Some of the approximations appear critical for short retrial intervals τ, particularly when c is small. The amount of effort and computer time spent on solving the retrial problem analytically is not necessarily less than the amount of effort and time spent by simulation. The problem is by no means solved, but it is hoped that by this analytical approach the way is paved toward a more complete analysis of retrial systems.

For message storage, the average delay can be determined by the well-established methods of traffic theory developed by A. K. Erlang. These are outlined in Appendix C. The average delay for the storage model is computed from (32) of Appendix C, and no approximations need to be made.

IV. RESULTS AND CONCLUSIONS

For a fixed amount of information load, each switching model produces different delays. This means that some switching models must be operated at lower occupancy than others to ensure that delays encountered will not exceed the desired maximum. The delays shown below for each switching model do not necessarily keep their relationships in respect to each other when messages are switched through several switching centers in tandem.

The results of computations for one address per message are shown in Figs. 2 and 3 for one and ten lines per group, respectively. Figs. 4 and 5 show similar results for three addresses per message and one and ten lines per group, respectively. The "information occupancy" in these figures is numerically equal to the information load offered to the line group divided by the number of lines per group, i.e., a_I/c. The term "occupancy" refers to the percentage of time a line is occupied on the average. The fraction of time a line is actually utilized for the transmission of information, then, is equal to a_I/c, so that we may also call "information occupancy" the "line utilization."

First let us discuss the address camp-on model. This method offers the advantage that error correction can be performed on multilink connections on an end-to-end basis because the message content remains in storage at the data source until a connection is set up to all addressed receiving stations. The camp-on model also is of interest because storage

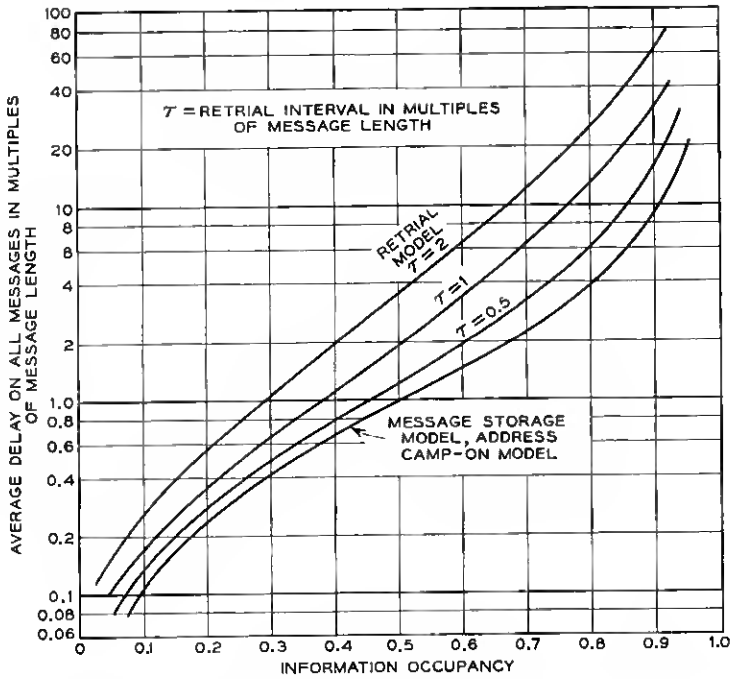

Fig. 2 — Average delay on all messages vs information occupancy; 1 line per group, 1 address per message.

at an intermediate switching center need be provided only for the address portion of a blocked message and this might have some economic advantages over other switching models. For one address per message there is, in theory, no difference in the traffic delay performance of the address camp-on and message storage models when the message is switched only once. It should be remembered that in the camp-on model an intermediate switching center keeps the incoming lines busy in excess of the message length for the duration of a delay which, for a given information load a_I, increases the actual load carried. Because we consider only single-switched messages, no account is taken here of this type of line loading.

As was mentioned before, the total holding time of an outgoing line in the address camp-on model is made up by the excess holding time, which is the time spent waiting for other addresses to find lines, and by the actual message length. In Fig. 6 we show the total occupancy a_T/c versus the actual information occupancy or line utilization a_I/c for three

addresses per message. We see that the total occupancy approaches 100 per cent at a surprisingly low information load. This is due to the fact that the excess holding time increases the load on the outgoing line groups, which in turn increases delays and thus brings about longer excess holding times. This makes the camp-on model unusable beyond certain intolerably low levels of line utilization. For instance, in the three-address case, line utilization must be limited to about 14 per cent or 60 per cent for line group sizes of $c = 1$ or $c = 10$, respectively. It can be seen in Fig. 6 that beyond this point the total occupancy blows up and with it the delay imposed on a message. A similar result was

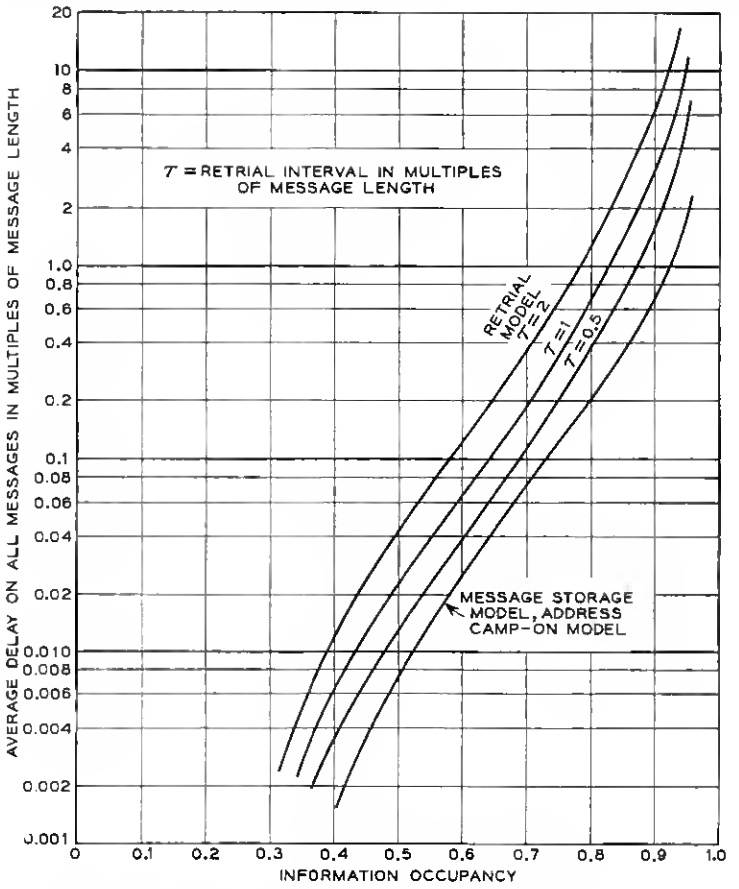

Fig. 3 — Average delay on all messages vs information occupancy; 10 lines per group, 1 address per message.

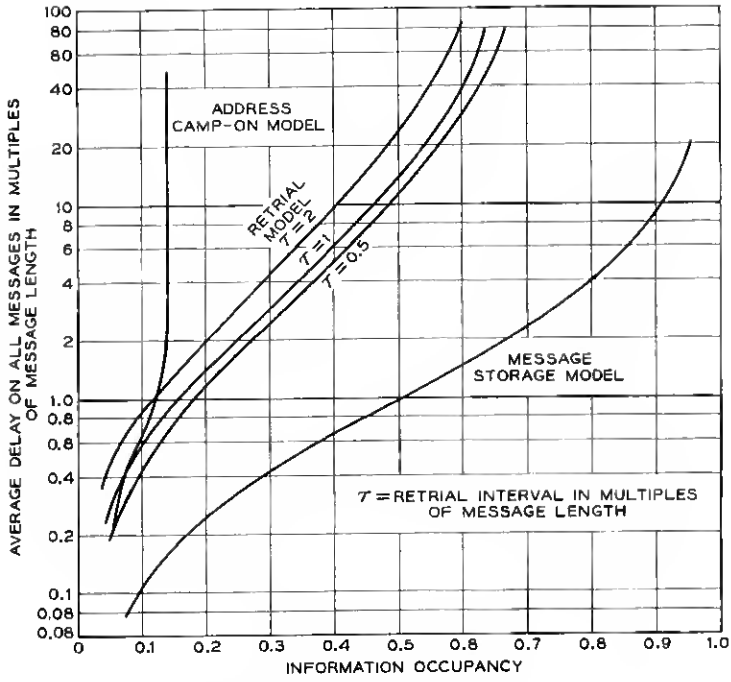

Fig. 4 — Average delay on all messages vs information occupancy; 1 line per group, 3 addresses per message.

obtained by Weber[1] in a different approach. A simulation made by Weber[1] for $c = 1$ shows close agreement with the results derived here. For example, the simulation shows a maximum utilization of about 14 per cent for $c = 1$ and three addresses per message, which is the same as derived here analytically. This indicates that considerable confidence may be placed in the approach presented in Appendix A.

From the delay performance of the address camp-on model it is concluded that any switching method in which delays become a substantial part of the line holding time will require a relatively large number of lines to provide adequate service. By the same token we may conclude that even more lines will be required when the message is switched more than once, i.e., through more than one switching center.

Next we turn our attention to the retrial model. The performance of the retrial model as a function of the retrial interval, τ, is of interest. We observe in Figs. 2–5 that when one doubles τ the delay is less than doubled. On the one hand, we expect longer retrial intervals to cause

longer delays. On the other hand, it can be shown that the probability that the message succeeds on a retrial increases with increasing length of τ. Shorter retrial intervals result in smaller chance for success than longer retrial intervals, but the fact that in any given time there are more attempts made with short retrials than with long retrials makes the average delay a monotone increasing function of τ. For large values

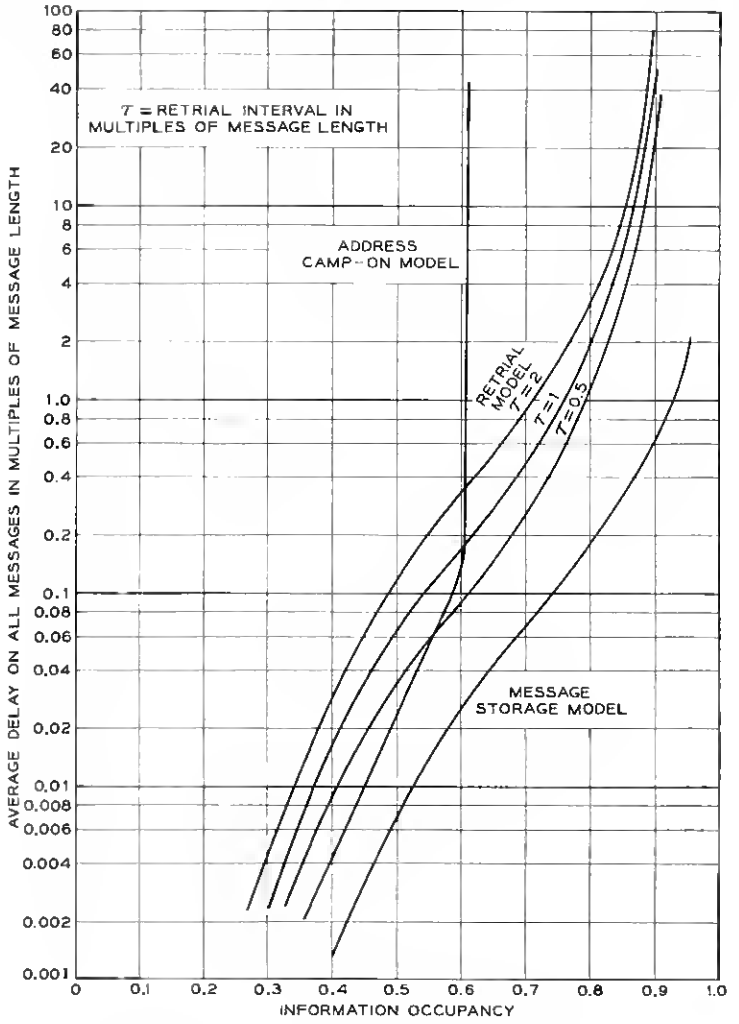

Fig. 5 — Average delay on all messages vs information occupancy; 10 lines per group, 3 addresses per message.

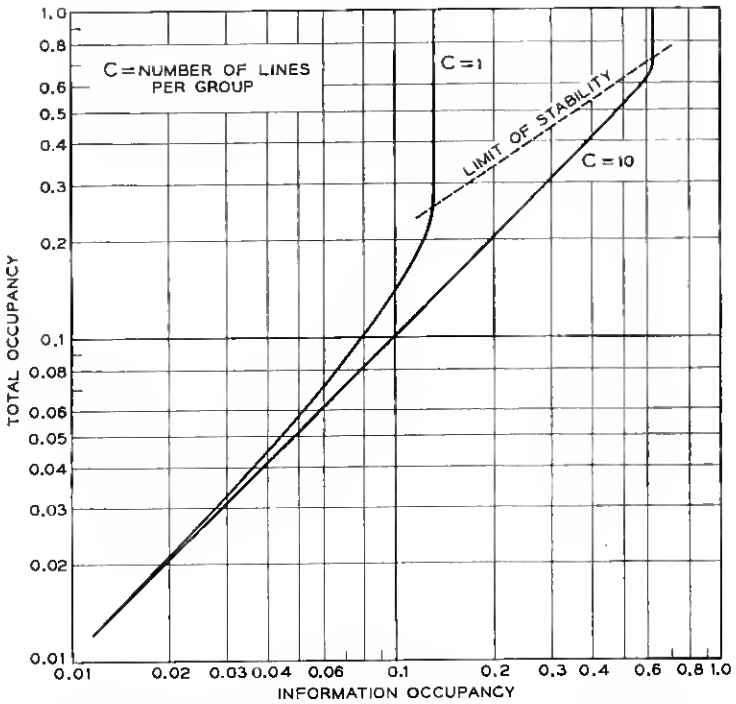

Fig. 6 — Total occupancy vs information occupancy for address camp-on model; 3 addresses per message.

of τ, say at least twice the message length, the rate of increase of the delay will be almost proportional to the rate of increase in τ. This is so because the rate of change of the probability of being blocked again becomes smaller with increasing values of τ. With τ approaching 0, the average delay with the retrial model should approach the delay for the message storage model in the case of one address per message.

The delay in the retrial model increases with the number of addresses per message. A comparison between Figs. 2 and 4 shows that this increase is quite significant when the number of lines per group is small. This increase can, of course, be avoided when a multiaddress message is broken up into several one-address messages, as suggested earlier in Section 2.3. Stations not blocked would then receive the message content independent of the availability of lines to the other stations. If this mode of operation is used for retrials, the retrial curves given in Figs. 4 and 5 are to be replaced by those for the one-address case shown in Figs. 2 and 3, respectively. Such a change of the retrial model would bring the av-

erage delay for very short retrial intervals close to that of the message storage model, independent of the number of addresses.

The retrial model has the advantage that no storage has to be provided in a switching center except at the data source. However, provision must be made to instruct the data source to reoffer the message when an all-lines-busy condition is encountered at a switching center. The retrial method is particularly well suited for error correction by retransmission from the data source on request from a terminating station which detected an error.

Since the analysis of the retrial model is based on approximations, we do not expect absolute accuracy of the curves derived. The retrial model has been simulated by others for some special cases and it is interesting to compare the results. This is done in Table I, in which time is expressed in units of the average message length. We observe some disagreement between analysis and simulation for large retrial intervals.

Finally, let us look at the message storage model. This method provides an efficient use of lines, even if the line groups are small, and delay is independent of the number of addresses per message. It requires, however, that considerable storage be provided because each switching center must be arranged to permit full message storage to allow for delays exceeding a message length. Provision must also be made for transmission of a copy of the message to each addressed station independent of the transmission to any other station.

According to the model of the message storage method, a message is delayed only until the very instant a line is found. From a practical point of view this means that the line-hunting mechanism should be activated as soon as the address is decoded. On the other hand, message storage may be operated so that the message is stored completely at the

TABLE I — COMPARISON: ANALYSIS VERSUS SIMULATION, ONE ADDRESS PER MESSAGE

Retrial Interval	No. of Lines	Occupancy a_1/c	Average Delay on All Messages		
			Analysis	Simulation 1	Simulation 2
0.5	1	0.7	3.25	*	3.65
0.5	10	0.7	0.11	*	0.14
1.0	1	0.5	1.94	1.84	1.76
1.0	10	0.9	3.12	2.61	1.92
2.0	1	0.7	12.65	*	8.45
2.0	10	0.7	0.36	*	0.26

* Not available.

switching center before the line hunting starts. In this ease, the amount of time needed for full message storage must be added to the delay. The latter mode of operation adversely affects the delay performance of the storage model, particularly when the message is long in comparison with the delay that can be tolerated.

The delay performance for the storage model will beeome considerably worse than shown when the line back to the originating station cannot be released as soon as the message has been transmitted over it. For multiswitched messages the release of lines between switching centers would ensure that the line holding time is not increased by the delay.

The curves given for the storage model can be considered accurate because the validity of Erlang delay formulas has long been observed.

V. ACKNOWLEDGMENTS

The author wishes to acknowledge the valuable assistance received from D. L. Clark, both in computor programming and in the mathematical derivations. Also acknowledged are suggestions made by E. Wolman in a critical review of this paper, and the encouragement received from E. E. Schwenzfeger during the preparation period.

APPENDIX A

Mathematical Analysis of Address Camp-On Model

Let the excess holding time E_A be the average time between seizure of a line and tbe time lines are found for A addresses of a given message. Further, let $Z_{K,A}$ be defined as the average time between initial request for lines and the time K out of A addresses have seized lines. At the time A addresses have seized lines, the message is ready to be transmitted. The average delay on all messages, d_A, is defined as the average time between initial request and tbe time lines have been seized by all A addresses, as illustrated in Fig. 7.

The average is a linear operator, and one obtains for the expeeted excess holding time

$$E_A = \frac{1}{A} \sum_{K=1}^{A} W_{K,A}.$$

And since

$$W_{K,A} = d_A - Z_{K,A}$$

it follows that

$$E_A = d_A - \frac{1}{A} \sum_{K=1}^{A} Z_{K,A}. \tag{1}$$

The term $\frac{1}{A} \sum_{K=1}^{A} Z_{K,A}$ is recognized as the average time between request for and seizure of a line for any given single address.

We define $Q(t)$ as the probability that the delay is less than or equal to t between the time of a request for a line by a given single address and the time of line seizure, and obtain

$$\frac{1}{A} \sum_{K=1}^{A} Z_{K,A} = \int_{t=0}^{\infty} t \, dQ(t).$$

The average excess holding time for the A-address case follows from (1) as

$$E_A = \int_{t=0}^{\infty} t \, d[Q(t)]^A - \int_{t=0}^{\infty} t \, dQ(t). \tag{2}$$

The only approximation in the analysis of the address camp-on model is the assumption that the holding time of a line is exponentially distributed, so that

$$Q(t) = 1 - \delta e^{-\varphi t}. \tag{3}$$

The holding time is made up of two random variables, namely the message length and the excess holding time. The approximation made in (3) implies that the sum of these two random variables is exponentially distributed. That this, indeed, is a reasonable assumption is confirmed

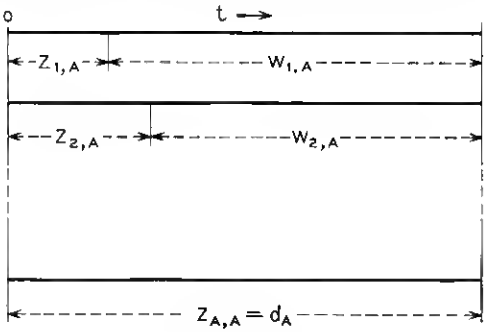

Fig. 7 — Line seizure sequence.

by the close agreement of the delay derived here with the delay derived by simulation.[1]

Substitution of (3) in (2) gives

$$E_A = A\delta\varphi \int_{t=0}^{\infty} e^{-\varphi t}\, t(1 - \varphi e^{-\varphi t})^{A-1}\, dt - \varphi\delta \int_{t=0}^{\infty} te^{-\varphi t}\, dt.$$

The above reduces with the binomial expansion for $(1 - \delta e^{-\varphi t})^{A-1}$ to

$$E_A = -\frac{A}{\varphi} \sum_{K=0}^{A-1} \binom{A-1}{K} \frac{(-\delta)^{K+1}}{(K+1)^2} - \frac{\delta}{\varphi}. \tag{4}$$

For the case of exponential line holding time with mean $\bar{t}$ and service of requests for lines in the order of arrival as is the case here, we must substitute in (4) according to Erlang[2]

$$\delta = 1 - Q(0) \tag{5}$$

$$\varphi = \frac{c - a_T}{\bar{t}} = \frac{c - a_T}{a_T/a_1}, \tag{6}$$

in which c is the number of lines per group, a_T the average number of requests per line holding time or the total load offered per line group, and a_I the average number of requests per message length or the information load. $Q(0)$, the probability of no delay, is given by Erlang[2] as

$$Q(0) = 1 - \frac{\dfrac{a_T{}^c e^{-a_T}}{c!}\,\dfrac{c}{c - a_T}}{1 - \displaystyle\sum_{i=c}^{\infty} \frac{a_T{}^i e^{-a_T}}{i!} + \frac{a_T{}^c e^{-a_T}}{c!}\,\frac{c}{c - a_T}}. \tag{7}$$

The unit of time being the average message length renders for $\bar{t}$, the average line holding time,

$$1 + E_A = \bar{t} \tag{8}$$

so that

$$E_A = \frac{a_T}{a_I} - 1. \tag{9}$$

Substitution of (5), (6) and (9) into (4) brings

$$\frac{a_T}{a_I} - 1 + \frac{a_T}{a_I}\,\frac{1 - Q(0)}{c - a_T} = -\frac{a_T}{a_I}\,\frac{A}{c - a_T} \cdot \sum_{K=0}^{A-1} \binom{A-1}{K} \frac{[Q(0) - 1]^{K+1}}{(K+1)^2}. \tag{10}$$

With c, a_I and A given, we can now compute a_T, the total load, itera-

tively from (7) and (10) above. Then, with a_T known, we find the average delay on all messages from (1) and (9) as

$$d_A = \frac{a_T}{a_I} - 1 + \frac{1 - Q(0)}{c - a_T} \frac{a_T}{a_I}, \tag{11}$$

which we recognize as the left-hand member of (10).

It is interesting to take note of the fact that there are generally two values of a_T which fulfill (10). Thus, one could conclude that the system can operate in two modes, one implying a shorter delay than the other. More than one steady state of operation has been observed by others[1,5] in similar traffic studies. However, it appears questionable that the heavy delay mode is stable since the larger of the two a_T which fulfill (10) decreases with increasing a_I, which is physically unreasonable.

APPENDIX B

Mathematical Analysis of Retrial Model

When a newly offered message makes its first trial to seize one of c lines in a group, let S_i, $0 \leqq i \leqq c$, be the unconditional probability that the group is in state i. A line group is said to be in state i when i out of all c lines in the group are busy. The message is reoffered until a line is available to each of A addressed receiving stations; therefore, no messages are lost. The load carried on each group equals a_I, the information load offered. For the special case of $c = 1$, we obtain

$$S_0 = 1 - a_I \tag{12}$$

and

$$S_1 = a_I. \tag{13}$$

For $c > 1$, i.e., for more than one line per group, S_i depends not only on a_I, but also on the procedure by which lines are made busy. By procedure is meant the type of distribution of the length of the intervals between line requests, and whether unsuccessful attempts form queues or are withdrawn. For Poisson input at the rate of a_0 and withdrawal of blocked attempts, Erlang loss probability[3] gives

$$S_i = \frac{a_0{}^i/i!}{\sum_{x=0}^{c} a_0{}^x/x!} \tag{14}$$

in which

$$a_0 = \frac{a_I}{1 - S_c}. \tag{15}$$

In the switching model considered here, the total input to each line group, i.e., the total load offered, is not Poisson distributed and its magnitude differs from a_0. This is so because repeated attempts are blocked with a probability which is larger than S_c and because for $A > 1$ blocking to any one of the A addressed receiving stations will cause a retrial. In order to compute S_i, it is assumed that for a sufficiently large retrial interval τ, the total input will at least resemble a_0 in distribution and magnitude. This approximation may be justified for light line occupancies but it becomes increasingly unrealistic with increasing line occupancies. The approximation for a_0, therefore, is used here only to compute values for S_i when $c \geqq 2$. For $c = 1$ the values for S_i are exactly determined by (12) and (13) above. For $c > 1$ we solve (14) and (15) iteratively with $i = c$ to obtain a_0 and then solve (14) to obtain approximations for S_i when $c \geqq 2$. The approximations for S_i so obtained are used for the unconditional state probability both in the one and three-address cases.

Next, we will consider conditional probabilities which take into account past known states of a line group. Let it be known that at a given time t_0 there are j lines busy in a group, $0 \leqq j \leqq c$; what then is the probability that at $t_0 + t$ there are i lines busy? This conditional probability is called $X_{i,j}(t)$. For Poisson input and exponential line holding time, $X_{i,j}(t)$ is given by a well-known set of first-order differential equations.[4] Here, now, we must take into account that the superposition of first and repeated attempts results in an input which is not Poisson. We let $\omega(t)$, $t_0 \leqq t \leqq t_0 + \tau$, be the instantaneous line request rate or the density of requests. As was said above, $X_{i,j}(t)$ is conditioned on state j of the group at t_0. Consequently $\omega(t)$ depends also on the state of the group at t_0, and this important point should be kept in mind, particularly since the notation does not always remind the reader of this condition.

Assume for the time being $\omega(t)$ is known for every value of t in the interval $(t_0, t_0 + \tau)$. The differential equations defining $X_{i,j}(t)$, $t_0 \leqq t \leqq t_0 + \tau$, are

$$
\begin{aligned}
X_{0,j}'(t) &= -\omega(t) \cdot X_{0,j}(t) + X_{1,j}(t) \\
X_{i,j}'(t) &= \omega(t) \cdot X_{i-1,j}(t) - [i + \omega(t)] \cdot X_{i,j}(t) + \\
&\qquad (i + 1) \cdot X_{i+1,j}(t) \qquad (16) \\
&\text{for } 0 < i < c \\
X_{c,j}'(t) &= \omega(t) \cdot X_{c-1,j}(t) - c \cdot X_{c,j}(t).
\end{aligned}
$$

The condition that j lines are busy at time t_0 is taken into account by

$$X_{i,j}(t_0) = \begin{cases} 1 \text{ for } i = j \\ 0 \text{ for } i \neq j, \end{cases} \tag{17}$$

and, as said before, by $\omega(t)$. For the special case $t = t_0 + \tau$, the argument is dropped and $X_{i,j}(t_0 + \tau)$ is abbreviated to $X_{i,j}$. The system (16) can be solved with Laplace transforms for $c \leqq 2$, and by numerical integration for $c \geqq 3$.

Two cases are considered in the following: one in which each message has one address, i.e., the case $A = 1$; the other in which each message has three addresses, i.e., the case $A = 3$.

B.1 *One Address per Message*

In order to compute delay for the case $A = 1$, we must know the probability of finding the line group in state c at $t_0 + \tau$, given a state c at t_0, $t_0 - \tau$, $t_0 - 2\tau$, etc. In other words, we must know the probability that a message is blocked twice, three times, four times, etc. To simplify the notations for the case $A = 1$, we write C_i for the conditional probability that all c lines of a group are busy at $t_0 + \tau$, given a state c at t_0, $t_0 - \tau, \cdots, t_0 - i\tau$. C_i, therefore, denotes the probability that a one-address message is blocked $i + 2$ times in a row. C_0 is identical with $X_{c,c}$ and can be computed by (16) provided $\omega(t)$ is known. It will help to keep matters clear if, for the case $A = 1$, $\omega(t)$ is subscripted so that $\omega_i(t)$ refers to the condition that at t_0, $t_0 - \tau, \cdots, t_0 - i\tau$ all c lines of the group are known to be busy. For instance, the line request rate used in (16) to compute C_0 is called $\omega_0(t)$. Values for C_i, $i \geqq 1$, are computed from (16) in the same manner as $X_{c,c}$, except that $\omega_i(t)$ is conditioned as indicated later. Hence, the numerical values for $X_{c,c}$ obtained from (16) with $\omega_i(t)$, $i \geqq 0$, are equal to C_i.

Let us now discuss the procedure by which $\omega_i(t)$ is obtained for the case $A = 1$. We will find functions L_i which are conditioned on a state c at t_0, $t_0 - \tau, \cdots, t_0 - i\tau$, such that

$$\omega_i(t) = L_i(C_0, C_1, C_2, \cdots, t). \tag{18}$$

This means that we cannot obtain an explicit expression for $\omega_i(t)$ since $\omega_i(t)$ is needed to compute C_i from (16). But with (18) we come into a position which allows us to assume values for C_i, compute $\omega_i(t)$ from (18) and then use the so-computed $\omega_i(t)$ to obtain C_i from (16). Through successive iterations stable solutions are obtained for C_i such

that (19) is satisfied

$$S_o < C_0 < C_1 < C_2 < \cdots < 1. \tag{19}$$

The above approach, of course, is extremely tedious when C_i must be computed for large numbers of i. The problem is simplified when one assumes for large enough i that $C_i = C_{i+1}$. Computations have shown that for retrial intervals $\tau \geqq 0.5$ we may reasonably approximate $C_i = C_{i+1}$ when $i \geqq 4$. This approximation, consequently, is used in the one-address case. It is mainly for this approximation that the analysis for the case $A = 1$ is limited to retrial intervals $\tau \geqq 0.5$.

Before we define the function L_i in (18), we will give the method by which the successive iterations are performed by a computer program to compute C_i. We start by iterating for a stable value of C_0 with $C_i = C_0$ for $i \geqq 1$. Next, we iterate for a stable value of C_1 with C_0 fixed and $C_i = C_1$ for $i \geqq 2$. Now we go back and iterate for a new value of C_0 with C_1 fixed and $C_i = C_1$ for $i \geqq 2$. This process is continued until no further changes in C_0 and C_1 are detected. Continuing one step further, we iterate for a stable value of C_2 with C_0 and C_1 fixed and $C_i = C_2$ for $i \geqq 3$. Again, we back up and search for a new value of C_0 with C_1 and C_2 fixed, then search for a new value of C_1 with C_0 and C_2 fixed and finally search for a new value of C_2 with C_0 and C_1 fixed, all with $C_i = C_2$ for $i \geqq 3$. We proceed in steps in the manner described above until finally no changes are detected in C_0, C_1, C_2, C_3, C_4 with $C_i = C_4$ for $i \geqq 5$.

With the approximation $C_i = C_4$ for $i \geqq 5$, we can write for (18)

$$\omega_i(t) = L_i(C_0, C_1, C_2, C_3, C_4, t). \tag{20}$$

Line requests are made by first and repeated attempts. First attempts arrive independent of time with a density v. Repeated attempts arrive with a density $u_i(t)$, in which i refers to the condition that all lines are busy at time t_0, $t_0 - \tau$, $\cdots$, $t_0 - i\tau$ and t is some time such that $t_0 \leqq t \leqq t_0 + \tau$. With these definitions we substitute for $L_i(C_0, C_1, C_2, C_3, C_4, t)$ in (20)

$$\omega_i(t) = v + u_i(t). \tag{21}$$

The density of first attempts, according to definition, is numerically equal to the information load offered or

$$v = a_I.$$

The density of repeated attempts in the interval $(t_0, t_0 + \tau)$ is derived from first attempts which are made before t_0 and are blocked. For in-

stance a kth attempt, $k \geqq 2$, occurs at t, $t_0 \leqq t \leqq t_0 + \tau$, if the attempt occurs first at $t - (k-1)\tau$ and all lines are busy at $t - (k-1)\tau$, $t-(k-2)\tau, \cdots, t-\tau$. As an abbreviation we write $\Pr(t_x | t_1, t_2, t_3, \cdots)$ for the probability that all lines are busy at t_x, conditioned on all lines busy at t_1 and t_2 and $t_3 \cdots$. As before, let t be an instant in time such that $t_0 \leqq t \leqq t_0 + \tau$. With the condition that all lines are busy at t_0, $t_0 - \tau, \cdots, t_0 - i\tau$, we obtain for the density of repeated attempts at t

$$\begin{aligned} u_i(t) = a_I[&\Pr(t - \tau \mid t_0, t_0 - \tau, \cdots, t_0 - i\tau) \\ &+ \Pr(t - 2\tau \mid t_0, t_0 - \tau, \cdots, t_0 - i\tau) \\ &\cdot \Pr(t - \tau \mid t_0, t_0 - \tau, \cdots, t_0 - i\tau, t - 2\tau) \\ &+ \Pr(t - 3\tau \mid t_0, t_0 - \tau, \cdots, t_0 - i\tau) \\ &\cdot \Pr(t - 2\tau \mid t_0, t_0 - \tau, \cdots, t_0 - i\tau, t - 3\tau) \\ &\cdot \Pr(t - \tau \mid t_0, t_0 - \tau, \cdots, t_0 - i\tau, t - 3\tau, t - 2\tau) + \cdots]. \end{aligned} \tag{22}$$

We are left with the problem of expressing $\Pr(t_x \mid t_1, t_2, t_3, \cdots)$ in the above as functions of C_0, C_1, C_2, C_3 and C_4. Assume that symmetry exists such that for any positive length of time l

$$\Pr(t_x + l \mid t_x, t_x - \tau, \cdots, t_x - k\tau) = \Pr(t_x - l \mid t_x, t_x + \tau, \cdots, t_x + k\tau).$$

In the above it is assumed that traffic congestion builds up to an all-lines-busy condition at t_x, $t_x + \tau, \cdots, t_x + k\tau$ in the same manner as it subsides after $t_x + k\tau$. This assumption may not be exact for the retrial system but this concept is used here since it is expected to give a good enough approximation for the following reason.

If a group is busy, say, at t_0, then it must be expected that part of the traffic which contributes to the congestion at t_0 is reoffered traffic. The fact that congestion occurs at t_0 implies that all lines were busy at $t_0 - \tau$, $t_0 - 2\tau$, etc., with a larger probability than indicated by the unconditional state probability S_c. As an approximation to the function by which traffic is expected to build up we construct linear functions in time. For example, we assume that the probability of blocking at some time $t_x < t_0$ builds up to an all-lines-busy condition at $t_0 - \tau$ and t_0 as shown in Fig. 8. Also, we assume independence of events that are not really independent. For instance, we assume that blocking between $t_0 - \tau$ and t_0 occurs with a probability $W_{1,1}(t)$ as defined below. Similarly, independence is assumed between events occurring with probability $N_i(t)$, $M_i(t)$ or $W_{i,j}(t)$ and the event which causes a repeated attempt

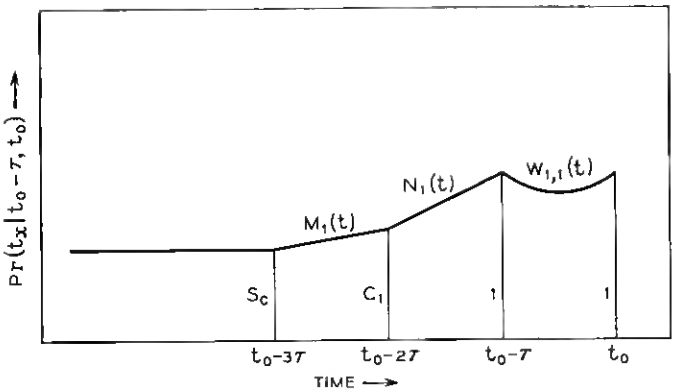

Fig. 8 — Sample of build-up function.

at some time prior to t_0. The approximations made in the expressions below may account for some of the differences which are observed between theory and simulation.

A first attempt, made at some time $t - j\tau$, $t_0 \leqq t \leqq t_0 + \tau$, $j \geqq 1$, is blocked in the nomenclature of (22) with a probability

$$\Pr\,(t - j\tau \mid t_0\,,\, t_0 - \tau,\, \cdots,\, t_0 - i\tau)$$

for which we approximate

$$N_i(t) = \frac{1}{\tau}\,[t + C_i(\tau - t)] \qquad \text{for } j = i + 1,\ i \geqq 0$$

$$M_i(t) = \frac{1}{\tau}\,[C_i t + S_c(\tau - t)] \qquad \text{for } j = i + 2,\ i \geqq 0$$

$$S_c \qquad \text{for } j \geqq i + 3,\ i \geqq 0$$

and with $N_i^*(t) = (1/\tau)[t(C_i - 1) + \tau]$

$$W_{i,j}(t) = 1 - [1 - N_{j-1}(t)][1 - N_{i-j}^*(t)]$$

$$\text{for } 1 \leqq j \leqq i,\ i \geqq 1.$$

A kth repeated attempt made at some time $t - j\tau$, $t_0 \leqq t \leqq t_0 + \tau$, $j \geqq 1$, is blocked in the nomenclature of (22) with a probability

$$\Pr\,[t - j\tau \mid t_0\,,\, t_0 - \tau,\, \cdots,\, t_0 - i\tau,$$
$$t - (j + 1)\tau,\, t - (j + 2)\tau,\, \cdots,\, t - (j + k)\tau]$$

for which we approximate

$$\begin{aligned} &1 - (1 - C_{k-1})\,[1 - N_i(t)] && \text{for } j = i + 1,\ i \geqq 0 \\ &1 - (1 - C_{k-1})\,[1 - M_i(t)] && \text{for } j = i + 2,\ i \geqq 0 \\ &C_{k-1} && \text{for } j \geqq i + 3,\ i \geqq 0 \end{aligned}$$

and finally

$$1 - (1 - C_{k-1})\,[1 - W_{i,j}(t)] \qquad \text{for } 1 \geqq j \geqq i,\ i \geqq 1.$$

We make use of the above expressions as shown in (22) to obtain $u_i(t)$. The subscript i of $u_i(t)$ corresponds with the value of i in the above approximations and refers to the condition that all lines are busy at t_0, $t_0 - \tau$, $\cdots$, $t_0 - i\tau$. For every such i, $i = 1, 2, 3$ and 4, we have, according to (21), a $L_i(C_0, C_1, C_2, C_3, C_4, t)$ and an $\omega_i(t)$, and can perform the iterations outlined before to compute C_0, C_1, C_2, C_3, and C_4.

Continuing in the analysis of the one-address case, we will now evaluate the delay. The probability that a one-address message is delayed exactly $i\tau$ is given by $D_1(i)$ which is

$$\begin{aligned} D_1(0) &= 1 - S_c \\ D_1(1) &= S_c(1 - C_0) \\ D_1(i) &= S_c(1 - C_{i-1}) \prod_{j=0}^{i-2} C_j, \quad i \geqq 2. \end{aligned}$$

$H_1(i)$, the probability that the delay is greater than $i\tau$ for the one-address case, is

$$H_1(i) = 1 - \sum_{k=0}^{i} D_1(k), \qquad i \geqq 0$$

which reduces to

$$\begin{aligned} H_1(0) &= S_c \\ H_1(i) &= S_c \prod_{j=0}^{i-1} C_j, \qquad i \geqq 1. \end{aligned}$$

The average delay is obtained as the summation of all possible delay values multiplied by their respective probability of occurrence and is given by

$$d_1 = \tau \sum_{i=1}^{\infty} i\, D_1(i)$$

which reduces to

$$d_1 = \tau S_a[1 + \sum_{i=0}^{\infty} \prod_{j=0}^{i} C_j]. \tag{23}$$

The above formula is used to compute the average delay on all messages for the case $A = 1$. The unconditional probability of finding all lines busy, represented by S_c, and the conditional blocking probabilities C_i are approximated by the methods outlined before.

B.2 *Three Addresses per Message*

The delay for the case $A = 3$ is computed in a Markov process. This means that we are considering only a first-order dependency, since we make the assumption that the conditional probability $X_{i,j}$ of finding i lines of a group busy at $t_0 + \tau$ depends only on state j of that group at t_0. This and the following approximations appear justifiable in the multi-address case when considering the multitude of factors which determine the line request rate between t_0 and $t_0 + \tau$. The principal assumption for the case $A = 3$ is that ω, the sum of the densities of line requests of first and repeated attempts during t_0 and $t_0 + \tau$, is nearly Poisson distributed and therefore independent of any states at or before t_0. Recall that for $A = 1$ we have been concerned only with the conditional probability of state c at $t_0 + \tau$ given also a state c at t_0 or at t_0 and $t_0 - \tau$, etc. For the case $A = 3$, however, we are concerned with an $X_{i,j}$ for all values of i and j, $0 \leqq i \leqq c$, $0 \leqq j \leqq c$, as will become apparent later. For a known ω, we obtain $X_{i,j}$ from (16). The condition that j trunks are busy, now, is accounted for only by the initial condition as given in (17). The density of line requests ω is obtained similarly to (21) as the sum of the densities of first attempts a_I, and of repeated attempts u which, according to our assumptions for $A = 3$, are time independent. It is obvious that in the case of A addresses per message, $A > 1$, a line request in a given group is made only when the condition is fulfilled that the remaining $A - 1$ groups are not busy. Since independence is assumed we can set $(1 - S_c)^2$ for this condition in the three-address case and obtain

$$\omega = (a_I + u)(1 - S_c)^2 \tag{24}$$

for the density of line requests in the interval $(t_0, t_0 + \tau)$ in any given line group. The expression given in (24) above, of course, is an approximation since in reality u is dependent on the state of the group at t_0 and since independence is assumed between the event causing a repeated

attempt during $(t_0, t_0 + \tau)$ and the event causing all lines to be busy in the other two addressed line groups. But, as said before, these dependencies are believed to be noncritical for the multiaddress case, so that ω is considered to be independent of time.

As in the one-address case, we are left with the problem of defining u, the density of repeated attempts, which is expressed below as a function of $X_{c,c}$. We obtain u by the following approach. Let G be the probability that a first attempt is blocked at some time prior to t_0. This probability is approximated by

$$G = 3S_c(1 - S_c)^2 + 3S_c^2(1 - S_c) + S_c^3. \tag{25}$$

The above is an approximation because it assumes independence between the event which causes a group to be in state j, $0 \leqq j \leqq c$, at t_0 and the event causing all lines of a group to be busy at some time prior to t_0. For the probability that a kth attempt, $k \geqq 2$, is blocked prior to t_0 we approximate

$$\begin{aligned} R = {} & X_{c,c}(1 - S_c)^2 + 2S_c(1 - X_{c,c})(1 - S_c) \\ & + 2X_{c,c}S_c(1 - S_c) + S_c^2(1 - X_{c,c}) + X_{c,c}S_c^2 \end{aligned} \tag{26}$$

for which it is assumed that at the $k - 1$st attempt one line group was in state c, i.e., busy, but without having made any assumptions about the state of the remaining two groups. The expression given for R in (26) is an approximation since, as before, the known state of a group at t_0 is ignored and since only a first-order dependency is considered, as mentioned earlier. The density of repeated attempts is obtained similarly to the one-address case by considering all attempts which were blocked prior to t_0 so that

$$u = a_I G + a_I G R + a_I G R^2 + \cdots$$

or

$$u = a_I G \frac{1}{1 - R}. \tag{27}$$

Substitution of (27) in (24) gives

$$\omega = a_I\left(1 + \frac{G}{1 - R}\right)(1 - S_c)^2. \tag{28}$$

R in (28) above is a function of $X_{c,c}$. This means that we cannot find ω explicitly since ω is needed to compute $X_{c,c}$ as outlined in (16). To find ω we again must iterate by assuming a value for $X_{c,c}$ in (26), recompute

$X_{c,c}$ from (16) and then use the recomputed $X_{c,c}$ in (26). After having found a stable value for ω we can, from (16), readily compute $X_{i,j}$ for all $0 \leqq i \leqq c, 0 \leqq j \leqq c$.

In order to compute the delay for the three-address case we consider $P_{i,j,h}(k)$, which is defined as the probability of finding the three groups in state i, j and h respectively at the k + 1st attempt, $k \geqq 0$. $P_{i,j,h}(k)$ is obtained recursively by finding all possible ways in which the states of the three groups have changed to states i, j and h at the k + 1st attempt when at least one group was busy at the kth attempt. Using the approximation of a first-order dependency, as mentioned before, we get for $k \geqq 1$

$$\begin{aligned} P_{i,j,h}(k) = &\sum_{r=0}^{c}\sum_{s=0}^{c} P_{c,r,s}(k-1)\cdot X_{i,c}\cdot X_{j,r}\cdot X_{h,s} \\ &+ \sum_{q=0}^{c-1}\sum_{s=0}^{c} P_{q,c,s}(k-1)\cdot X_{i,q}\cdot X_{j,c}\cdot X_{h,s} \\ &+ \sum_{q=0}^{c-1}\sum_{r=0}^{c-1} P_{q,r,c}(k-1)\cdot X_{i,q}\cdot X_{j,r}\cdot X_{h,c}\,. \end{aligned} \tag{29}$$

At the first attempt,

$$P_{i,j,h}(0) = S_iS_jS_h\,.$$

A three-address message succeeds at the kth attempt, $k \geqq 0$, when at the kth attempt all three groups are in states other than c. The probability of a delay of exactly $k\tau$, then, is given for the three-address case by

$$D_3(k) = \sum_{i=0}^{c-1}\sum_{j=0}^{c-1}\sum_{h=0}^{c-1} P_{i,j,h}(k). \tag{30}$$

$H_3(k)$, the probability that the delay is greater than $k\tau$ for the three-address case, is

$$H_3(k) = 1 - \sum_{i=0}^{k} D_3(i), \qquad k \geqq 0.$$

As in the one-address case, we find the average delay on all messages for the three-address case by summing over all possible delay values multiplied by their respective probability of occurrence

$$d_3 = \tau \sum_{k=1}^{\infty} k\cdot D_3(k)$$

or with (30)

$$d_3 = \tau \sum_{k=1}^{\infty} \sum_{i=0}^{c-1} \sum_{j=0}^{c-1} \sum_{h=0}^{c-1} k \cdot P_{i,j,h}(k). \tag{31}$$

Recall that $P_{i,j,h}(k)$ is obtained reeursively as sbown in (29). The unconditional state probability S_i and the conditional probability $X_{i,j}$ which are both needed in (29) are approximated as described earlier.

APPENDIX C

Mathematical Analysis of Message Storage Model

The delay for the message storage model is eomputed by well-known methods of traffic theory and is given here only for reasons of completeness. The delay in the delivery of a eopy of the message to a given station is — according to the switching model — independent of the delay in the delivery to any other station. Delayed messages form queues in the order of arrival. An analysis for queued service and exponential line holding time was made by A. K. Erlang.

According to Erlang[2] we find for the average delay on all messages to any given receiving station

$$d = F(0) \cdot \frac{1}{c - a_I} \tag{32}$$

with

$$F(0) = \frac{\dfrac{a_I^c e^{-a_I}}{c!} \dfrac{c}{c - a_I}}{1 - \displaystyle\sum_{i=c}^{\infty} \frac{a_I^i e^{-a_I}}{i!} + \frac{a_I^c e^{-a_I}}{c!} \frac{c}{c - a_I}}.$$

The delay distribution, expressed as the probability that the delay is greater than t, is computed from

$$F(t) = F(0) \cdot e^{-(c - a_I)t}.$$

The curves for the message storage model are caleulated from (32). One should, however, bear in mind that in eertain speeialized applications of data communieation a copy of the message must sometimes have been delivered to all addressed reeeiving stations before the message is of use to any one station. One would then be interested in tbe average delay until a line is found to the receiving station with the longest delay of all

stations addressed by the message. For A addresses per message this delay is given by

$$d_A = \int_0^\infty t\, d[1 - F(t)]^A,$$

which for $A = 3$ reduces to

$$d_3 = \frac{F(0)}{c - a_I}\left\{3 - \frac{3}{2}F(0) + \frac{[F(0)]^2}{2}\right\}.$$

REFERENCES

1. Weber, J. H., unpublished work.
2. Molina, E. C., Application of the Theory of Probability to Telephone Trunking Problems, B.S.T.J., **6**, July, 1927, p. 469 and p. 471.
3. Wilkinson, R. I., Theories for Toll Traffic Engineering in the U.S.A., B.S.T.J., **35**, March, 1956, p. 426.
4. Brockmeyer, E., Halstrom, H. L., and Jensen, A., The Life and Works of A. K. Erlang, Transactions of the Danish Academy of Technical Sciences, No. 2, 1948, p. 33.
5. Helly, W., unpublished work.